*Published online 10 May 2010* *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W545–W549***
*doi:10.1093/nar/gkq366*

# Dali server: conservation mapping in 3D

**Liisa Holm[1,2,*] and Päivi Rosenström[1]**

[1]Institute of Biotechnology and [2]Department of Biosciences, University of Helsinki, Helsinki, Finland

Received January 31, 2010; Revised April 12, 2010; Accepted April 24, 2010

## ABSTRACT

**Our web site (http://ekhidna.biocenter.helsinki.fi/dali_server) runs the Dali program for protein structure comparison. The web site consists of three parts: (i) the Dali server compares newly solved structures against structures in the Protein Data Bank (PDB), (ii) the Dali database allows browsing precomputed structural neighbourhoods and (iii) the pairwise comparison generates suboptimal alignments for a pair of structures. Each part has its own query form and a common format for the results page. The inputs are either PDB identifiers or novel structures uploaded by the user. The results pages are hyperlinked to aid interactive analysis. The web interface is simple and easy to use. The key purpose of interactive analysis is to check whether conserved residues line up in multiple structural alignments and how conserved residues and ligands cluster together in multiple structure superimpositions. In favourable cases, protein structure comparison can lead to evolutionary discoveries not detected by sequence analysis.**

## INTRODUCTION

Comparative analyses of protein sequences and structures play a fundamental role in understanding proteins and their functions. Assuming an evolutionary continuity of structure and function, describing the structural similarity relationships between protein structures allows scientists to infer the functions of newly discovered proteins.

The most widespread purpose of structural alignment has been to identify homologous residues (encoded by the same codon in the genome of a common ancestor). Mutations manifest in plastic deformations, shifts and rotations of the secondary structure elements (SSEs). A wide spectrum of structural alignment methods exist, which differ in their treatment of structural variations, scoring functions and optimization algorithms [reviewed in (1)].

We are aware of half a dozen web servers (2–7) that provide structure comparisons against the current, weekly updated Protein Data Bank (PDB) (Supplementary Table S1). Each server is unique because they employ different structure comparison methods. The Dali server has been around for 15 years in various incarnations (2,8). We have now implemented interactive visualization of user's structures to the Helsinki server, and present validation data for an improved database search protocol in DaliLite v.3.3.

The primary result of the database search is the list of structural neighbours and their corresponding structural alignment. Usually homologous proteins have stronger structural similarity than convergent folds. However, the amount of structural similarity (we use *Z*-scores) at the transition between divergent and convergent folds is family specific. Therefore, manual inspection of the results is recommended. Residue conservation is a particularly powerful means of highlighting which are the key residues in the structure, and so can usually help to pick out the most likely location of the protein's functional site(s) (9–11).

## DESCRIPTION OF THE SERVER

### Inputs

The input to the server is one or two protein structures in PDB format. The query structure can be specified as a PDB identifier plus chain identifier, or a PDB file uploaded by the user. There are three cross-linked query forms for the Dali server, Dali Database and pairwise comparison, respectively. For example, the entry point to the Dali server is http://ekhidna.biocenter.helsinki.fi/dali_server.

All backbone atoms (N, CA, C, O) are required and the minimum chain length is 30 amino acids. Backbone atoms may be reconstructed from a CA trace using the MaxSprout server at http://www.ebi.ac.uk/maxsprout.

External links to the Dali database should use http://ekhidna.biocenter.helsinki.fi/dali/daliquery?pdbid=1nnn&chainid=A, where 1nnn represents a PDB identifier and chainid is optional. Meta-servers may link to http://ekhidna.biocenter.helsinki.fi/dali/daliquery_txt?pdbid=1nnn&chainid=A, which directly returns the match list and alignment data as plain text.

*To whom correspondence should be addressed. Tel: +358 9 19159115; Fax: +358 9 19159079; Email: liisa.holm@helsinki.fi

## Processing

Queries to the Dali Database and pairwise comparison are processed interactively; the result is usually returned within a minute. The Dali server processes up to eight PDB searches in parallel, others are queued. Most PDB-search queries are processed in less than an hour. Results are stored on the server for two weeks. The results of identical queries are retrieved instantly from cache.

The Dali server and Dali database return only the best match of the query to each PDB structure. The pairwise comparison returns also suboptimal matches. The pairwise comparison is based on a systematic branch-and-bound search that returns non-overlapping solutions in decreasing order of alignment score (12). Suboptimal matches can be of interest in cases of internal symmetries or repeated domains.

Dali Database is updated twice a year and contains precomputed structural alignments of PDB90 against the full PDB. The query structure is mapped to the closest representative in PDB90 and the structure comparison scores are recomputed using the transitive alignment via the representative.

The Dali server aims to retrieve a list of 500 structural neighbors of the query structure with the highest $Z$-scores (see Mathematical Appendix in Supplementary Data). Most query structures have strong similarity to a structure already in the PDB. We use fast filters to identify a shortlist of about 100 promising candidates (2). If these produce strong matches, the search proceeds by walking. Otherwise, the query structure is compared with PDB90 in one versus all fashion, followed by a walk to collect matches to redundant PDB structures (which are over 90% sequence identical to PDB90 representatives).

Walking selects targets for structural comparison from the neighbours of neighbours found so far (Figure 1). The second shell of neighbors is known because all structures in the PDB are stored in a precomputed network of similarities. The pairwise alignments (Q,P) and (P,R) induce a transitive alignment (Q,R), which is used as the starting point of refinement rather than optimizing the alignment from scratch. There are many possible choices of intermediate structure P en route from Q to R. We select the 'high road', in other words, the minimum of the $Z$-scores $Z$(Q,P) and $Z$(P,R) should be as high as possible. The 'high road' may change as more structures are added to the first neighbour shell. To avoid redundant comparisons, we only test induced alignments which are longer than previously obtained ones. When the alignment (Q,R) has been refined, R is added to the first neighbour shell. The walk ends when either there are no new neighbours in the second shell, a specified number of hits (1000) have been reported, or a maximum number of comparisons (1000) have been performed.

## Outputs

The Dali server, Dali Database and pairwise comparison use a common output format and share interactive analysis tools.

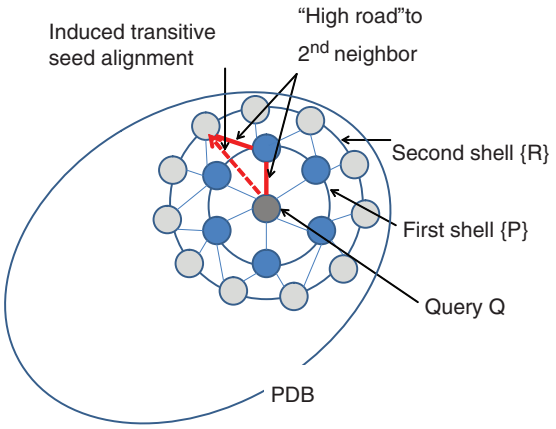


**Figure 1.** The walking strategy focuses the search for structural neighbours to the neighbourhood of similar structures found so far. A sparse network of alignments within PDB is stored in an internal database. Once query Q is aligned to some PDB structure(s), a complete set of similar structures can be collected by walking, provided that the network is connected and similarity is approximately transitive.

The result consists of (i) a list of structural neighbours, ranked by $Z$-score, and (ii) the alignment data. The results are presented as plain text for downloading by downstream application, and as hypertext for interactive analysis. The default results page reports the top 500 matches to all chains in the PDB. A subset of matches to PDB90, filtered at 90% sequence identity, is provided for convenience.

Selected subsets of matches can be visualized (i) as multiple sequence alignments, or (ii) in multiple 3D superimposition. While sophisticated tools with integrated sequence alignment and structure superimposition views are available, we have chosen Jmol, an open source Java viewer (http://www.jmol.org) for molecular graphics, because it was most easily accessible to the casual user. Each neighbour is aligned (superimposed) against the query structure in a star-like tree topology. Active sites can be recognized by clusters of conserved residues and ligands. Sequence and structure conservation are calculated within the selected subset of matches. Sequence conservation is given by Equation (1):

$$h(i) = \sum_{j=1}^{20} p(i,j)\, log_2 \left( \frac{p(i,j)}{q(j)} \right), \quad (1)$$

where $i$ is a position, $j$ represents the 20 natural amino acid types, $p$ is the frequency of an amino acid type in position $i$ of the alignment and $q$ is its frequency in the database.

Structure conservation is given by Equation (2):

$$s(i) = \frac{1}{N} \sum_{k=1}^{N} \partial(k,i), \quad (2)$$

where $k$ represents the structural neighbours, $N$ is the number of selected structural neighbours, and $\partial$ is 1 if position $i$ of structure $k$ is aligned to the query structure and 0 otherwise.

### Downloads

The Dali database and the DaliLite software are available for academic use from http://ekhidna.biocenter.helsinki.fi/dali/downloads/download.html.

## RESULTS

### Benchmarking

Newly solved protein structures are compared against those in the PDB in the hope of discovering remote evolutionary connections. Homologous proteins should rank at the top of the match list. The ability of the walking strategy to retrieve biologically interesting hits was assessed by comparison with the SCOP classification (13), which is a hierarchical classification of protein structures and curated by experts. The area under the coverage-reliability plot (AUC) was used for assessment. The average AUC per query is 0.79, 0.87 and 0.92 at fold, superfamily and family level, respectively (Figure 2).

A number of factors influence this evaluation. First, it has been shown that ranking by *Z*-scores as in Dali, or by other measures of structural similarity, approximates but does not reproduce the scop hierarchy at any cut-offs; the scop classification considers also other criteria than structural similarity (14). Secondly, the sampling of a structural neighbourhood during walking is dependent on the connectivity of the underlying (sparse) network of precomputed structural similarities. Thirdly, our measures of structural similarity are not metric, which means that matches cannot be retrieved in strictly decreasing order of *Z*-scores. Fourthly, optimization from the seed alignment may or may not converge to the global optimum. Finally, relaxed criteria are often used to account for inconsistencies and possible classification errors in SCOP; for example, pairs with different superfamily but identical fold would be excluded from the evaluation of superfamily recognition (15). Here, the 'true positive' set consists of all pairs with identical SCOP classification and the 'false positive' set contains all other pairs above a given *Z*-score in the match list. The evaluation thus gives an underestimate of accuracy.

### Example

As an example, we have selected a putative bacterial cell invasion protein (PDB entry 3kk7). The protein contains a membrane-attack complex/perforin (MACPF) domain. The same fold is utilized for defence in the immune system and for attack in bacterial cholesterol-dependent cytolysins (CDCs). The three key pieces of evidence for homology are functional similarity (pore formation), conservation of three glycine residues at a hinge and conservation of the complex core fold (16–18).

Interactive analysis (Figure 3) starts with eyeballing the match list. Dali retrieves the entire superfamily. Contact map-based methods such as Dali are tolerant of structural plasticity and typically generate longer alignments than methods based on root mean square deviation (RMSD) criteria. The extent of the common core is clearly seen when structural conservation is mapped onto the query structure. Conserved residues can be identified in the multiple sequence alignment as well as highlighted in the 3D structure.

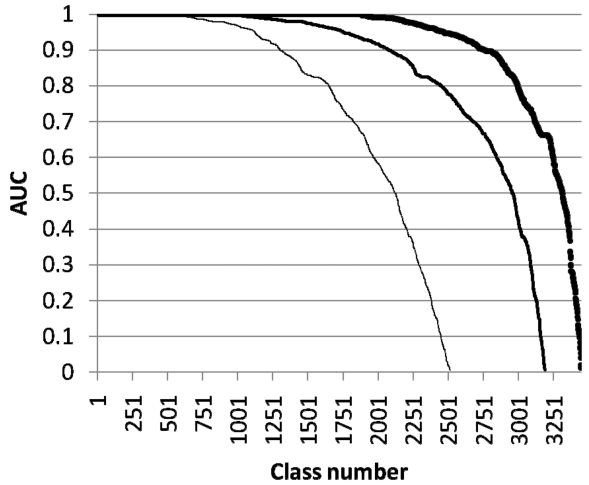


**Figure 2.** Benchmarking the walking strategy. The plots show the distribution of AUC values for single-domain queries in scop 1.75. AUC, where coverage is TP/T and reliability is TP/P with T the number of members of a scop class, P number of matches above a given *Z*-score, and TP the number of members of a scop class above a given *Z*-score. Evaluated classes are scop folds (thin line), scop superfamilies (medium thick line) and scop families (thick line). Classes with fewer than 10 or more than 500 members are excluded. Query structures consisted of single domains from scop classes a-d, filtered at 90% sequence identity. Matches to all PDB structures which contained an instance of a domain in scop classes a-d were included in the evaluation.

## CONCLUSIONS

In favourable cases, structural neighbours give clues of molecular function. The Dali server performs the search and provides convenient visualization tools to map conservation in 3D.

The Dali server has evolved to cope with the growth of the PDB, which is now a hundred times bigger than when the server started (8). Hundreds of new entries are added to the PDB every week. One versus all comparison is not affordable and overly aggressive pruning of search space may lose interesting matches. We think that the Dali server is operating with a satisfactory balance between speed and sensitivity.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## FUNDING

Funding from Academy of Finland project 109849. Funding for open access charge: Biocenter Finland

*Conflict of interest statement*. None declared.

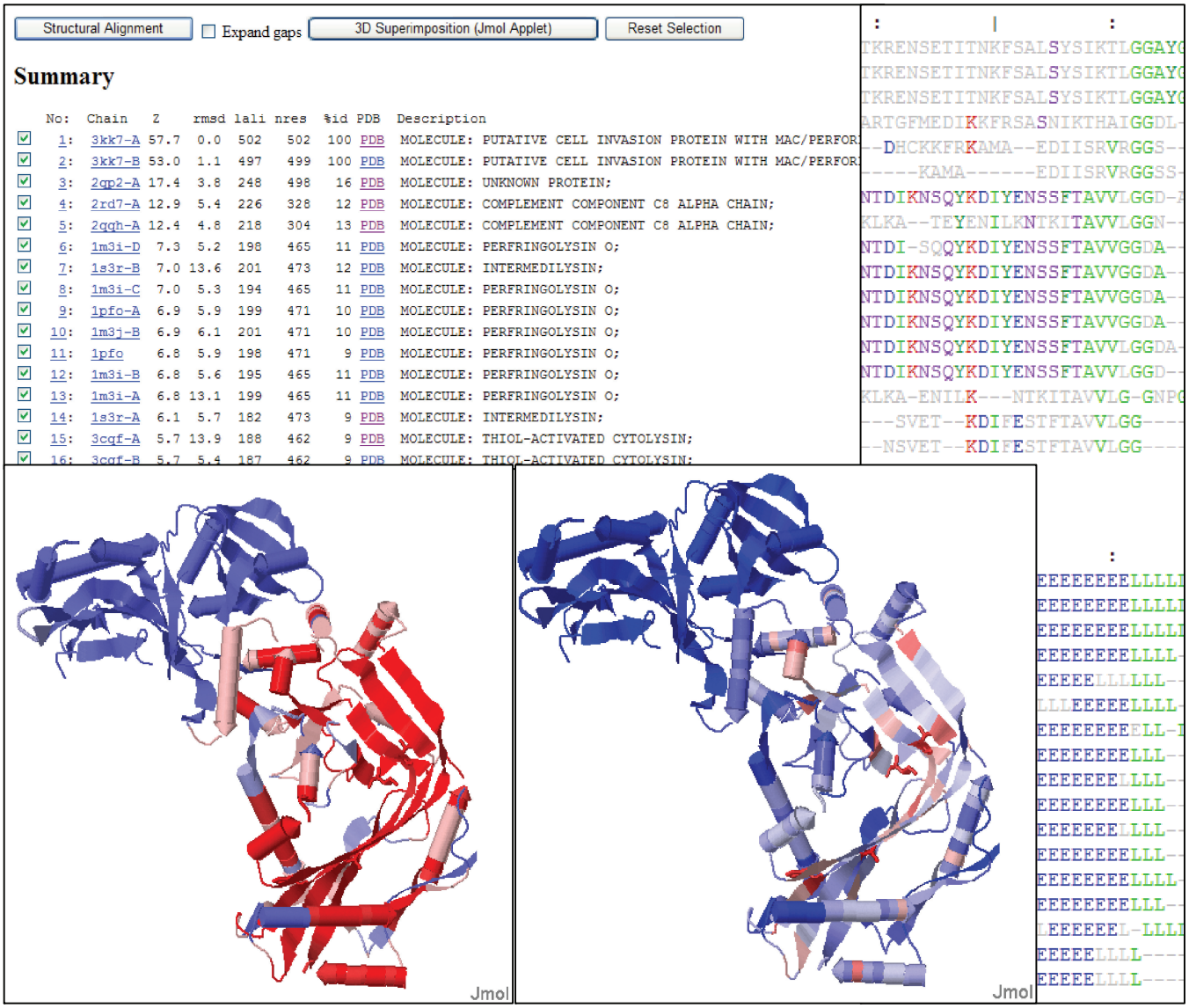


**Figure 3.** Using 3kk7A as query, the Dali server produces several views of the result, including: match list (top left), sequence view (top right), secondary structure view (bottom right), mapping of structural conservation (bottom left) and mapping of sequence conservation (middle bottom). The match list is sorted by *Z*-scores. The sequence view uses colour for amino acids which have a frequency above 0.5 in a column. The secondary structure view displays the alignment of helices (H), strands (E) and loops (L) as defined by DSSP. Sequence and structural conservation is computed within the selected subset of matches; red positions are maximally conserved and blue positions are not conserved.

## REFERENCES


1. Hasegawa,H. and Holm,L. (2009) Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.*, **19**, 381–389.
2. Holm,L., Kääriäinen,S., Rosenstrom,P. and Schenkel,A. (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.
3. Veeramalai,M., Ye,Y.Z. and Godzik,A. (2008) TOPS++FATCAT: Fast flexible structural alignment using constraints derived from TOPS+ string model. *BMC Bioinformatics*, **9**, Art. No. 358.
4. Kawabata,T. and Nishikawa,K. (2000) Protein tertiary structure comparison using the Markov transition model of evolution. *Proteins*, **41**, 108–122.
5. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst.*, **D60**, 2256–2268.
6. Gibrat,J.-F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
7. Leslin,C.M., Abyzov,A. and Ilyin,V.A. (2007) TOPOFIT-DB, a database of protein structural alignments based on the TOPOFIT method. *Nucleic Acids Res.*, **35**, D317–D321.
8. Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.

9. Lichtarge,O. and Sowa,M.E. (2002) Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.*, **12**, 21–27.
10. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, W299–W302.
11. Laskowski,R.A., Watson,J.D. and Thornton,J.M. (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **33**, W89–W93.
12. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
13. Andreeva,A., Howorth,D., Chandonia,J.-M., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2007) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
14. Sam,V., Tai,C.H., Garnier,J., Gibrat,J.F., Lee,B. and Munson,P.J. (2006) ROC and confusion analysis of structure comparison methods identify the main causes of divergence from manual protein classification. *BMC Bioinformatics*, **7**, 206.
15. Lindahl,E. and Elofsson,A. (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.*, **295**, 613–625.
16. Rosado,C.J., Buckle,A.M., Law,R.H., Butcher,R.E., Kan,W.T., Bird,C.H., Ung,K., Browne,K.A., Baran,K., Bashtannyk-Puhalovich,T.A. *et al.* (2007) A common fold mediates vertebrate defense and bacterial attack. *Science*, **317**, 1548–1551.
17. Hadders,M.A., Beringer,D.X. and Gros,P. (2007) Structure of C8alpha-MACPF reveals mechanism of membrane attack in complement immune defense. *Science*, **317**, 1552–1554.
18. Lukoyanova,N. and Saibil,H.R. (2008) Friend or foe: the same fold for attack and defense. *Trends Immunol.*, **29**, 51–53.